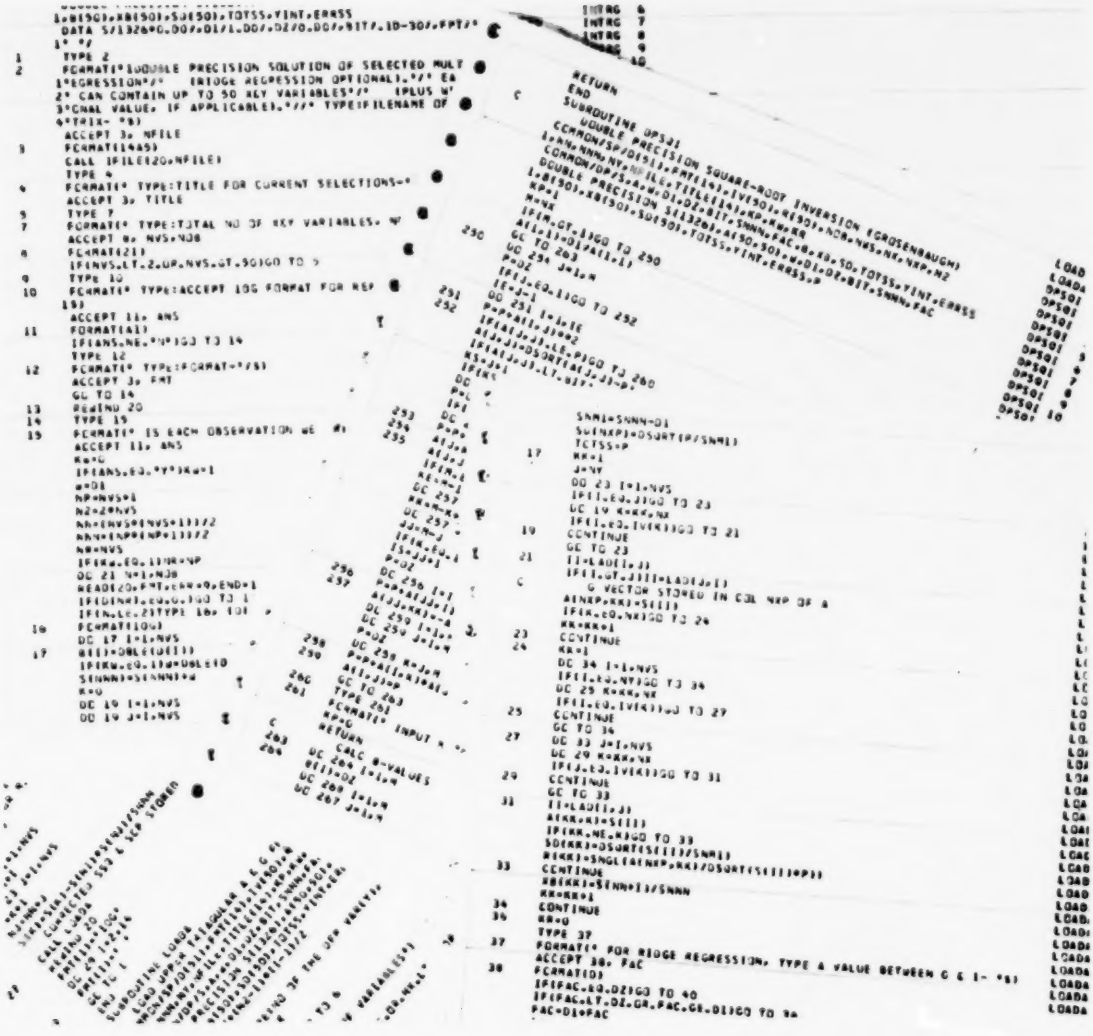
Forest Research Report No 106

INREG

An Interactive Computer Program for Multiple Linear Regression by Least Squares

F. L. Raymond



Ministry of Natural Resources
Ontario

Hon. James A. C. Auld
Minister

Dr. J. K. Reynolds
Deputy Minister

Forest Research Report No. 106

1980

ISSN 0301-3924

INREG

An Interactive Computer Program for Multiple Linear Regression by Least Squares

by

F.L. Raymond

Ontario Forest Research Centre
Ministry of Natural Resources
Maple, Ontario L0J 1E0

Ontario Ministry of Natural Resources

Hon. James A. C. Auld
Minister

Dr. J. K. Reynolds
Deputy Minister

RAYMOND, F.L. 1980. INREG--An interactive computer program for multiple linear regression by least squares. Ont. Min. Natur. Resour., Forest Res. Rep. No. 106. 21 p.

A programming package written in FORTRAN IV is made available for performing multiple linear least-squares regression in an interactive mode via a computer terminal. Its use should facilitate understanding and modeling the underlying relationships among many kinds of descriptive variables. Weighting of observations, to correct for nonhomogeneity of variance or for regressions on means of grouped data, and biasing by ridge regression, to correct signs of least-squares coefficients, are provided as options.

INTRODUCTION

Regression is an extremely complex subject and its frontiers are still being extended. Many of us who wish to apply regression tend to treat it as a not-very-well-understood tool that is made available at a computer installation. Many such computer-based tools encourage us to lose contact with the data and discourage us from improving our understanding of those tools and their limitations. This report provides an alternative tool which encourages maintenance of contact with the data; it also represents a modest attempt at making some regression problems more readily understood.

Problems in applying regression are almost separable into four categories: (1) erratic data, (2) collinearity, or ill conditioning of the data matrices, such as is caused by interdependencies among the "independent" variables, (3) nonhomogeneity of the variance of the dependent variable, and (4) selecting the best regression equation and interpreting the results it provides. This report will discuss mainly the latter two.

Daniel and Wood (1971) address the problems with erratic data by examining residuals (deviations between observed and estimated values) obtained by familiar procedures. Belsley *et al.* (1980) deal extensively with identifying and evaluating the influence of problems with either data or collinearity. However, they utilize totally different calculation procedures for accomplishing this and for arriving at the same least-squares equations. Those procedures appear very promising but they are generally outside the scope of the present paper. Belsley *et al.* reported that scaling all of the observations for each independent variable so that each such subset of observations totals to unity can substantially reduce collinearity. Scaling requires further study. They pointed out that each particular scaling has diverse effects at various stages in their calculations. Furthermore, when Lawson and Hanson (1974) utilized many of the same procedures, they apparently preferred that the scaling should cause each subset of observations to total Euclidean unity, not unity.

Regarding nonhomogeneity of variance, some users accomplish approximate homogeneity by transforming the dependent variable before performing the regression. This is not recommended here because it causes the resulting equation to be a biased estimator of that variable in its original units, unless the amount of bias is known and corrected for. Apparently such corrections are known only for square-root and logarithmic transformations. The recommendation here is for accomplishing approximate

homogeneity of variance by weighting the observations after the manner of Cunia (1964).

Selection of any regression equation depends upon the purpose for which it is to be used. No single equation can be "best" for all purposes. Draper and Smith (1966) identified the dilemma facing regression users as follows: (1) To make the equation useful for predictive purposes we want to include as many independent variables as possible. (2) To minimize the costs of measuring independent variables for subsequent predictions we want to include as few as possible.

Each particular resolution of that dilemma can be viewed as a matter determined by how much precision of estimation is required for the particular purpose. If the original set of data is representative and covers the full ranges to be found in subsequent measurements of the variables, then every least-squares equation can be an unbiased estimator and the precision of estimation tends to increase with the number of independent variables in the equation.

If the purpose is to obtain an equation for use in making a very large number of estimates of, say, standing timber volume[1], then one wants to minimize subsequent measurement costs and usually is prepared to sacrifice precision by using a small equation. There are many least-squares programming packages available that do an efficient job of finding such small equations for the user. It matters very little that, as pointed out by Draper and Smith (1966), no two packages will necessarily select the same subset of variables.

However, if the purpose is either to be certain that the equation will be relatively unbiased in all locations or to obtain an understanding of the functional relationships in order to model them and to forecast future growth and yields, then one wants to begin with a large number of variables. For example, there are preliminary indications that geographic location may explain appreciable amounts of variation in the timber volume on plots of standing trees in boreal Ontario. If this proves to be the case, then any equation not large enough to include variables related to geographic location would give biased estimates of volume in most parts of that region, even though the combined estimate for the whole of boreal Ontario might be totally unbiased. Clearly, regression analysis should not always be treated as a one-shot, automatic process providing a small equation. An appropriately small equation can always be derived from a large one.

Furthermore, the selection criterion utilized by these popular packages tends to be related to maximization of the square of the multiple correlation coefficient (RSQ)[2]. Grosenbaugh (1967) pointed out that the residual, or error, mean square (ERRMS) is a more useful selection criterion when one is interested in precision. However, ERRMS does not lend

---

[1]Forestry terms are used throughout the text; however, other biologists will be able to recognize analogies to their own problems. The examples given in Appendices B and C are essentially chemical.

[2]Acronyms such as RSQ and ERRMS are the labels used in the Fortran programs.

itself to providing a convenient "stopping rule". His answer to the automatic stopping provided by maximizing RSQ was his extremely versatile programming package REX, which examined every possible equation for any set of independent variables. The "best" equation was the one with the smallest ERRMS.

Now, if there are p independent variables, there are 2^p-1 combinations of them, or equations. Because of the cost of computing all possible combinations, and especially of computing the larger combinations, Grosenbaugh suggested limiting p to 17.

When it is considered that, in addition to the usual variables describing the stands or the trees of any species, one might want to include others describing physiography, geography, stand structure, interspecific competition and, if possible, stand history, then it is clear that limiting p to 17 can be far too restrictive.

REX and most of the automatic small-equation packages were designed for use in a batch mode of processing. This requires the user to submit his data for subjection to a series of computations; he cannot see, or interact with, any of the results until the package determines them to have been completed.

Fortunately, most of the larger present-day computing systems have since been designed or adapted to support an interactive mode of processing. This permits the user to specify a trial equation via a computer terminal, to observe the results of the computations almost immediately and to specify the next trial. If one begins by including a very large number of variables, deletes several of least significant of them, and periodically attempts to reinsert many of the deleted variables, one will shortly arrive at the equation with the smallest ERRMS, or at something very close to it. This "put and take" procedure applied in an interactive mode can help the user find such equations more inexpensively than is possible by computing all of the combinations when p is large.

DISCUSSION

In "The analysis and selection of variables in linear regression", Hocking (1976) reviewed basic concepts, computational techniques, selection criteria and biased estimation.

Under the heading "Selection Criteria", he pointed out the importance of relating the selection of equations to their intended uses. He then discussed the criteria. He noted that most users prefer RSQ as a selection criterion, while many statisticians prefer ERRMS. He discussed several other candidate criteria, all of which appear to be elaborations of RSQ, ERRMS or of the residual, or error, sum of squares (ERRSS).

Because least squares is based upon minimization of the ERRSS, one might wonder why that value is rarely discussed for equation selection. Grosenbaugh (1967) rejected both ERRSS and RSQ as selection criteria on the basis that they increased or decreased monotonically as the number of variables increased. Furthermore, ERRMS is the basis for determining the precision with which any equation fits the data from which it was derived.

The standard error of the estimate (SEE) for any equation is the square root of ERRMS; the ratio between the mean of the dependent variable and SEE appears to be the best indicator of such precision.

It may be difficult for some to accept the equation with the minimum ERRSQ and SEE as a "best" equation in those cases where some of the variables it contains have t-values that are less than significant. This may result from interdependencies among the theoretically independent variables. A t-value for any regression coefficient indicates significance of the independent contribution made by that variable to regression as expressed in the context of the particular equation. Some less-than-significant variables seem to be making an additional joint, or supportive, contribution to other variables. That is, the presence of an independently nonsignificant variable often appears to reinforce the contributions of other variables. If deletion of a nonsignificant variable has a marked impact on the SEE, the ERRMS and on the t-values of the remaining variables, then its contribution is nontrivial and it should be left in the equation during the early stages of regression analyses for the purpose of understanding relationships. If anyone cannot accept a "best" equation with some of the t-values less than significant, or believes others will not, then it is only necessary to delete the least significant variables until a near-best equation is attained with all remaining variables making significant individual contributions[3].

Regarding biased estimation, Hocking (1976) listed three classes of biased "alternatives" to least-squares regression. These were: (1) Bayesian maximum-likelihood regression, (2) principal components regression and (3) ridge regression. Only the first of these is a true alternative; the others are adjuncts to least-squares regression.

Hocking cited Stein shrinkage as employing maximum likelihood estimation under the assumptions that the dependent and independent variables are multivariate normal and that their means are known. Neither of those conditions is ever likely to be true for forestry data.

Williams (1959) also discussed the use of maximum likelihood estimation in cases where the variance of the dependent variable is some given function of its expected value. He felt that doing so made use of the information added by estimating the variance. However, he pointed out that the quantity of such information is generally small and may be ignored; for this reason, and for computational simplicity, weighted least-squares regression is generally used instead of maximum likelihood regression, according to Williams.

I am inclined to believe that the act of fitting data to any given or easily specified mathematical function tends to distort the interpretation of those data at least a little. Maximum likelihood procedures appear to be extensively based on just such "fitting" of data to predetermined functions. Consequently, I prefer least-squares regression because it appears to approximate letting the data determine which equation fits them the best, and to cause the smallest distortions.

---

[3]The possibility that one or more "significant" variables may have a spurious relationship cannot be ignored.

Principal components regression amounts to converting the interdependent variables into an equal number of mutually independent components by calculating the eigenvalues. It makes the assumption that the number of variables to be deleted in least-squares regression equals the number of components with near-zero eigenvalues. Any components with near-zero eigenvalues may be considered trivial in eigenvector space, but a large act of faith is required to assume that an equal number of interdependent variables will be trivial in variable space. Furthermore, eigenvalues as indicators of component importance are clearly sensitive to changes in the scales used to measure or report the variables.

Ridge regression is the biased "alternative" in which Hocking (1976) expressed the greatest interest. He cited Hoerl and Kennard (1970) as having established the existence of a ridging constant that provides better estimates of the regression coefficients than can be obtained by least squares. Whether this is true or not outside of the theoretical case, it is true that a small amount of ridging will sometimes correct the signs of regression coefficients when they differ from those of the corresponding correlation coefficients.

Ridge regression amounts to increasing every element of the main diagonal of the sums-of-squares-and-cross-products (SSCP) matrix, or of the correlation matrix, by a fixed percentage or ridging constant before inversion. Hocking noted that much of the discussion of ridge regression is focussed upon the choice of the ridging constant. Mitchell and Hann (1979) produced their programming package BREX (for biased REX) to batch-process ridge regressions applying a series of trial constants to any best least-squares equation. They reasoned that a biased estimator which is close to the true mean with high probability, is better than an unbiased estimator which has a low probability of being near the true value. Their solution to the problem of selecting the best ridging value was to choose the one which made the SSCP matrix behave as though all of the original independent variables had been normal and totally uncorrelated.

That solution appears to be a matter of carrying a good thing too far. Forcing interdependent, non-normal variables to behave as though they were uncorrelated and normal is drastic treatment. Furthermore, every increase in the ridging constant apparently causes a corresponding increase in the ERRMS and the SEE; this must tend to offset any "high probability" there may have been for keeping the estimates near the mean.

Belsley *et al.* (1980) considered the possibility of utilizing the Bayesian procedures as a means of correcting for collinearity by incorporating prior information. However, they rejected this in favor of a "mixed-estimation" procedure. In this, any prior information in the form of expected relationships among the variables was simply added directly to the data matrix. Their second choice, especially where comparatively objective prior information was unavailable, was ridge regression.

The option of ridging any weighted or unweighted least-squares equation is included in INREG. For anyone wishing to explore the effects of mild biasing by ridging, the recommended procedure is to use ridging constants small enough so that there is little inflation of the SEE or the ERRMS, and to observe the effects upon the residuals, or deviations.

The option for permitting weighting of observations may be more useful in forestry applications. Data used to regress tree volume (Y) on other variables typically exhibit considerable heterogeneity of the variance (s^2) of the volume, with s^2 increasing dramatically as the volume or the variable for, say, diameter-breast-height squared times tree height, increases. This violates a fundamental least-squares assumption that the variance be homogeneous. Cunia (1964) discussed this situation in detail and gave several examples estimating s^2 by regressing it on functions of such a variable. He also illustrated and tabulated estimates of s^2 made by reading them from a hand-drawn curve; this appears to be the best and most practical method in many cases. Fitting such curves by regression seems an unnecessary and a possibly unsatisfactory complication.

To obtain useful estimates of s^2, order the observations by the magnitude of Y and partition them into groups, then calculate the variance of Y for each group and plot it against the group mean for Y; if the number of groups is small, then repartition and recalculate to obtain additional points; draw a smooth curve through the points and read from it the estimated value of s^2 for each observed Y. The appropriate weight for each observation is $1/s^2$ where s^2 is the estimate read from the curve. In theory, weights should be calculated, tried and recalculated; however, Cunia and Williams agree that the first weights obtained are rarely much improved upon by recalculation.

If each "observation" to be fitted is the mean of several original observations, then each point to be plotted is the s^2 of Y obtained from that subset of those original observations contributing to a given mean. The appropriate weight for each "observation" is n/s^2 where n is the number of original observations contributing to that mean. If all the means have essentially the same variance, then n can be an appropriate weight; however, n/s^2 is the weighting factor preferred for use with INREG.

The fitting of unweighted observations requiring weighting often gives equations for which the largest percentage deviations from regression tend to be found among those observations with the smallest Y-values. As a result, curves expected to pass through, or near, the origin will not do so. This may not be a bad situation where one is most interested in estimating the volume of the largest trees; however, to use such equations as a basis for modeling growth would be misleading. If those observations are weighted as suggested above and then fitted, the percentage deviations will be more homogeneous and more evenly distributed over the range of Y-values, and the curves will fit better near the origin.

Regardless of how the regression is performed, problems of interpreting the results are likely to exist. Sometimes these relate to decisions about the appropriate treatment of outlying observations; that is, ones which are aberrant even after weighting. Daniel and Wood (1971) provide insight into this.

More frequently, the problem is one of making sense of the signs and magnitudes of particular regression coefficients. Every equation treats a universe that is influenced by the variables it contains and by no others. Every regression coefficient may have been adjusted by the act of fitting to compensate for the mutual contributions of other variables. This can sometimes be resolved by considering only related subsets of variables rather by

trying to consider one variable in isolation. For example, when an equation contains D, D^2, H and D^2H, some of their regression coefficients may be negative even though each has a positive simple correlation with the dependent variable. Related variables should be considered jointly. It may sometimes be useful to regress the dependent variable on only a problem subset. However, less distortion may be introduced if one first does a fit without such variables, calculates the deviations from regression and then fits those deviations to the subset of problem variables. This amounts to stagewise regression and may be useful even though it is not a true least-squares solution (*cf.* Draper and Smith 1966).

On the one hand, differences in signs are sometimes resolved or "corrected" by mild ridging. On the other hand, although ridging can correct the sign of the regression coefficient, weighting sometimes has the effect of "correcting" the sign of the simple correlation coefficient to match that of the regression coefficient.

Next, a comment about the internal evaluation of qualitative variables. For example, soil moisture regime usually is indexed into a series of numeric classes. Estimates of differences among classes can be calculated by adding "dummy" variables to the regression equation in the manner described by Draper and Smith (1966, p. 134-137). This requires one less dummy variable than there are classes. It is important that the regression equation used be close to a best equation, as previously defined, so that one is working with something approaching the "full model" of many authors. The particular variable should be removed and replaced by the set of dummy variables.

Finally, most large computing systems provide interactive regression packages. Unfortunately, those which use single-precision arithmetic appear accurate for small equations but estimate regression coefficients very poorly for large equations. Those which use double-precision arithmetic tend to be expensive and inconvenient to use, especially when they tend to reject any near-singular trials[4] without providing information as to the reason. INREG generally accepts such trials. The resulting t-values are inflated, but their ranking appears correct, so that one knows which variable to delete for the next trial. In cases of extreme near-singularity, INREG replaces the R- and RSQ-values with asterisks.

PROGRAM AVAILABILITY

INREG is written in FORTRAN IV and is operational for interactive use on a Digital Equipment Corporation PDP-10. The source programs are available as a card deck. Card decks and program listings may be obtained by writing to:

---

[4] This happens when there are so many variables and so much collinearity that RSQ approaches unity. Belsley *et al.* (1980) suggest that, as a general rule, p should not exceed 40% of the number of observations.

PROGRAM INREG
Ontario Forest Research Centre
Ministry of Natural Resources
Maple, Ontario, Canada L0J 1E0

PROGRAM FEATURES

In its present version, INREG requires 13K words of core memory. Hollerith words are five characters long in a PDP-10. To adapt the program to a system with four- or six-character words, the dimensioning and use of vectors FMT and TITLE will have to be changed. The $ which appears internally in all of the prompting questions is the conventional PAUSE (for a response) on a PDP-10. The file identification conventions will also differ from system to system; the file of observations is identified with Logical Unit 20 in the program. Because INREG is written in FORTRAN IV, it can easily be modified for use on other computing systems.

Many of the features, such as single-precision reading of data, double-precision compilation and storage in a "permanent" vector, and the loading of a "temporary" upper triangular matrix for double-precision square-root inversion, were borrowed directly from Grosenbaugh's REX. The ridging option is similar to Mitchell and Hann's BREX. They converted each SSCP matrix into its corresponding correlation matrix and inverted that to obtain standard partial regression coefficients[5]; however, since this appears not to increase computing precision, INREG inverts the SSCP matrix directly. This gives ordinary partial linear regression coefficients (b_i) that are biased or unbiased, depending on whether the ridging option was exercised or not.

Operation of INREG is controlled by a series of prompts. Each time a prompting message is issued, the system waits for a response followed by a carriage return. The expected responses are Y for yes, N for no, or one or more numbers, terminated by the carriage return key. If no number or letter is given before the carriage return, the program will interpret that action in whatever way permits the cheapest recovery from mistakes made when responding. In most instances, any inappropriate or out-of-range value will cause recycling back to one of the previously issued prompts.

In an attempt to make the program relatively efficient to use, the answering of prompting messages is kept as compact as feasible, and certain descriptors usually found in regression programs, such as the standard error of each regression coefficient[6] and the array of all possible variances and covariances, are not provided. An analysis of variance (ANOVA) table and a table of residuals are available as options at the end of each trial run.

---

[5] The standard partial regression coefficient for X_i = b_i * standard deviation of X_i/standard deviation of Y. *E.g.* for the first X in the first example in Appendix C, the standard partial regression coefficient is 1.46831 * 5.8824/15.044 = 0.57413.

[6] The standard error of the partial regression coefficient discussed in footnote 5 is b_i/t_i = 1.46831/12.1047 = 0.1213.

Values for the multiple correlation coefficient R, RSQ, SEE, ERRMS and ERRSS are given for every trial run. It should be noted that if the run involves weighted observations or ridge regression or both, these values and the others in the ANOVA table will not be in original units. To obtain SEE-, ERRMS- and ERRSS-values in original units in such cases, simply request a table of residuals.

SUMMARY

INREG does not pretend to solve all of the problems of regression; I agree with Cunia (1964) that some form of regression is needed which would allow for the fact that neither the dependent nor the independent variables can be measured without error. However, until such a form of regression has been developed and thoroughly tested, multiple linear least-squares regression appears to be the best we have for many purposes.

INREG makes least-squares regression a readily available and conveniently useable tool for anyone wishing to gain and utilize a better understanding of functional relationships among variables. It is not intended as a replacement for the many programming packages that are efficient at finding small equations where neither great precision nor understanding is required.

It is a programming package intended primarily for multiple linear least-squares regression, in an interactive mode. It borrows extensively from Grosenbaugh's REX. All of the important calculations are done in double-precision arithmetic. It permits weighting of observations both to counteract the tendency for the variance of any dependent variable to increase with its magnitude and to allow for regressions on class means[7]. It also supports ridge regression[8] for anyone wishing to explore that kind of biasing of any given, weighted or unweighted, least-squares equation.

Regression is a useful first step in modeling but the equations obtained may not be good models, especially for growth and yield where the models must be dynamic and must allow for extreme diversity about general trends. Mathematical models can provide elegant ways of describing those trends, but the simpler and more elegant they are, the more difficult it is to make real data fit them. Programming models, based on relationships explored through regression and designed to deal with diversity, may prove capable of following the development of individual stands more faithfully and with less distortion of reality than is possible with mathematical models, unless the latter are extremely complex.

---

[7] This latter case is illustrated in Appendix B.

[8] Examples of ridge regression are given in Appendix C.

ACKNOWLEDGEMENT

The writer wishes to extend his sincere thanks to: foresters H. Anderson, B. MacDonald, J. Mervart, J. Osborn and D. Weingartner and biologists D. Fraser, A. McCombie and B. Shuter, for reviewing the manuscript and making helpful suggestions. Most will see that many of their suggestions have been used; none can be blamed for any left unused.

REFERENCES

BELSLEY, D.A., E. KUH and R.E. WELSCH. 1980. Regression diagnostics: identifying influencial data and sources of collinearity. John Wiley and Sons, Inc., N.Y. 292 p.

CUNIA, T. 1964. Weighted least squares method and construction of volume tables. For. Sci. 10:180-191.

DANIEL, C. and F.S. WOOD. 1971. Fitting equations to data. Computer analysis of multifactor data for scientists and engineers. Wiley-Interscience, N.Y. 342 p.

DRAPER, N.R. and H. SMITH. 1966. Applied regression analysis. John Wiley and Sons, Inc., N.Y. 407 p.

GROSENBAUGH, L.R. 1967. REX--Fortran-4 system for combinatorial screening or conventional analysis of multivariate regressions. USDA For. Serv. Res. Pap. PSW-44. 47 p.

HOCKING, R.R. 1976. The analysis and selection of variables in linear regression. Biometrics 32:1-49.

HOERL, A.E. and R.W. KENNARD. 1970. Ridge regression: biased estimation for non-orthogonal problems. Technometrics 12:55-67. (Not seen, cited by Hocking 1976)

LAWSON, C.L. and R.J. HANSON. 1974. Solving least squares problems. Prentice-Hall, Inc., N.J. 340 p.

MITCHELL, B.R. and D.W. HANN. 1979. BREX: A computer program for applying ridge regression techniques to multiple linear regression. USDA For. Serv. Techn. Rep. INT-51. 25 p.

WILLIAMS, E.J. 1959. Regression analysis. John Wiley and Sons, Inc., N.Y. 214 p.

APPENDIX A

Program Usage Notes

INREG will accept observations with up to 50 dependent and independent variables, which may appear in any order, from the data file or matrix. If each observation is weighted, then the weight must be the last "variable"; it can be the 51st.

1. When it asks for the file name, the name can be from one to five characters in length. (In the PDP-10, the name must also have the extension .DAT).

2. When it asks that a title be given for the current set of regression runs, the title can be up to 70 characters in length.

3. Next it asks for the number of variables (weight excluded) and the number of observations. It expects two numbers with a space or a comma between them.

4. The built-in format for reading the observation file expects the values for each observation to be 10 to a line, and that there be a space or a comma between each pair of values. If your data do not suit these conditions, answer the question with an N (for no) and it will ask you to provide a suitable format. The maximum permitted length is 70 characters; opening and closing parentheses are not required.

5. If your observations are weighted, then respond to this question with a Y and a carriage return.

At this point it will display all the values for the first two observations.

6. It then asks for a number identifying the dependent variable and for the total number of independent variables (space or comma between).

7. It will then ask you to type that many numbers, in order (space or comma between), identifying those independent variables. If there are more than 20, press the carriage-return key after the 20th and continue typing the rest. If there are more than 40, the same action must be taken after the 40th.

If you responded at step 6 with a larger or smaller number of independent variables than intended, typing 999 and the appropriate number of carriage returns will cause that prompt to be reissued.

8. The prompt about ridging will accept several responses. A carriage return alone or a zero will cause ridging to be skipped; a value between zero and one will cause ridging with that value as the constant; any other value will cause the prompt to be reissued.

It will then display the fitted equation as a table of variable means, standard deviations, simple-correlation and partial-regression coefficients and t-values. These will be followed by a line giving values

for R, RSQ, SEE, ERRMS and ERRSS.

9. It will then ask if you want an ANOVA table.

10. It will ask if you want a table of residuals.

If you obtained residuals and if the run was with either weighted observations or ridge regression, the SEE-, ERRMS- and ERRSS-values will be displayed in original units.

11. At the end of each run or trial, it will ask if you wish another run with the same file. If your response is N (for no), it will go to step 1; otherwise it will go to step 6.

12. Termination of the runs is accomplished by aborting them. On a PDP-10, this is done by pressing the C key twice while holding down the CONTROL key.

APPENDIX B

Example of Weighted Observations given by
Daniel and Wood (1971, p. 306)

The following page illustrates the use of INREG as applied to a simple example of weighted observations. There were four observations in it, of which the second was a mean of four initial observations while the remainder were single observations.

This shows that R, RSQ, SEE, ERRMS, ERRSS and other ANOVA items are given in weighted units whereas the residuals, the SEE, ERRMS and ERRSS obtained by summing the squared residuals, are in original units.

```
DOUBLE PRECISION SOLUTION OF SELECTED MULTIPLE LINEAR REGRESSION
  (RIDGE REGRESSION OPTIONAL).
EACH OBSERVATION CAN CONTAIN UP TO 50 X&Y VARIABLES
  (PLUS WEIGHT AS ADDITIONAL VALUE, IF APPLICABLE).

TYPE:FILENAME OF OBSERVATION MATRIX- DANDW

TYPE:TITLE FOR CURRENT SELECTIONS-
WEIGHTED EXAMPLE GIVEN BY DANIEL & WOOD (1971), P.306

TYPE:TOTAL NO OF X&Y VARIABLES, NO OF OBS- 2,4

TYPE:ACCEPT 10G FORMAT FOR READING OBSERVATIONS(Y/N)?- Y

IS EACH OBSERVATION WEIGHTED(Y/N)?- Y

  10.00000        33.80000        1.000000
  20.00000        62.20000        4.000000

TYPE:NO OF THE DEP VAR(Y), NO OF INDEP VARS(X)- 2,1

TYPE:  1 NOS (IN ORDER) FOR INDEP VARS, MAX 20 PER LINE-
1

FOR RIDGE REGRESSION, TYPE A VALUE BETWEEN 0 & 1-


 7-MAY-80 FILE DANDW
WEIGHTED EXAMPLE GIVEN BY DANIEL & WOOD (1971), P.306
X#      MEAN          ST DEV        R(XY)          REG CO       T VALUE
 1 0.22857E+02 0.95119E+01 0.99983E+00 0.29657QE+01  76.3336
Y#
 2 0.71000E+02 0.28215E+02                0.32105QE+01

R=.99983 RSQ=.99966 SEE=.90525E+00 ERRMS=.81947E+00 ERRSS=.16389E+01

DO YOU WANT AN ANOVA TABLE(Y/N)?- Y

ANOVA TABLE
  SOURCE        DF          SS              MS              F
REGRESSION      1.    0.477492E+04    0.477492E+04  5826.8144
ERROR           2.    0.163895E+01    0.819474E+00
TOTAL           3.    0.477656E+04

DO YOU WANT A TABLE OF RESIDUALS(Y/N)?- Y

OBS#    OBSERVED Y   ESTIMATED Y    OBSY-ESTY   %(O-E)/E
   1  0.33800E+02  0.32868E+02  0.93158E+00    2.834
   2  0.62200E+02  0.62526E+02 -0.32632E+00   -0.522
   3  0.92000E+02  0.92184E+02 -0.18421E+00   -0.200
   4  0.12240E+03  0.12184E+03  0.55789E+00    0.458

IN ORIGINAL UNITS: SEE=.81225E+00 ERRMS=.65975E+00 ERRSS=.13195E+01
```

APPENDIX C

Examples of Least-Squares and Ridge Regression on the Hald Data given by Draper and Smith (1966, p. 178)

The Hald data, consisting of 13 observations on five variables, were utilized extensively by Draper and Smith in comparisons between the all-equations procedure and several small-equations procedures. After considering all of the possible equations, they preferred the regression of Y (variable 5) on 1 and 2, basing their judgment on improvements in RSQ, which they favor. It was the one picked by the backward elimination and the stepwise procedures.

Judgment based on sequential F-testing picked the regression of Y on 1, 2 and 4. It was also picked by the forward-selection procedure. Judgment based on minimum ERRSS would pick the same regression, with Y on 1, 2 and 3 as a very close second choice.

Mitchell and Hann (1979) also chose the Hald data to illustrate ridging. They ridged the full equation (Y on 1, 2, 3 and 4), presumably because it had the largest RSQ value.

The following pages give six runs with INREG on the Hald data. They illustrate least-squares regressions of Y on 1 and 2; Y on 1, 2 and 3; Y on 1, 2 and 4 and Y on 1, 2, 3 and 4; they also illustrate ridge regression of Y on 1, 2, 3 and 4 using the ridging constants: 0.03, which Mitchell and Hann preferred; and 0.002 which also corrected the sign of the regression coefficient for variable 3.

It is worth pointing out that, if these had been forestry data, we likely would have added some interactions, or cross-products, between variables and some transforms, such as squares or cubes, of individual variables.

```
DOUBLE PRECISION SOLUTION OF SELECTED MULTIPLE LINEAR REGRESSION
  (RIDGE REGRESSION OPTIONAL).
EACH OBSERVATION CAN CONTAIN UP TO 50 X&Y VARIABLES
  (PLUS WEIGHT AS ADDITIONAL VALUE, IF APPLICABLE).

TYPE:FILENAME OF OBSERVATION MATRIX- THALD

TYPE:TITLE FOR CURRENT SELECTIONS-
HALD DATA GIVEN BY DRAPER & SMITH (1966), P.178

TYPE:TOTAL NO OF X&Y VARIABLES, NO OF OBS- 5,13

TYPE:ACCEPT 10G FORMAT FOR READING OBSERVATIONS(Y/N)?- Y

IS EACH OBSERVATION WEIGHTED(Y/N)?-

  7.000000        26.00000        6.000000        60.00000        78.50000

  1.000000        29.00000        15.00000        52.00000        74.30000


TYPE:NO OF THE DEP VAR(Y), NO OF INDEP VARS(X)- 5,2

TYPE:  2 NOS (IN ORDER) FOR INDEP VARS, MAX 20 PER LINE-
1,2

FOR RIDGE REGRESSION, TYPE A VALUE BETWEEN 0 & 1-


 7-MAY-80 FILE THALD
HALD DATA GIVEN BY DRAPER & SMITH (1966), P.178
X#      MEAN          ST DEV        R(XY)          REG CO       T VALUE
 1 0.74615E+01 0.58824E+01 0.73072E+00 0.146831E+01   12.1047
 2 0.48154E+02 0.15561E+02 0.81625E+00 0.662250E+00   14.4424
Y#
 5 0.95423E+02 0.15044E+02                0.525773E+02

R=.98928 RSQ=.97868 SEE=.24063E+01 ERRMS=.57904E+01 ERRSS=.57904E+02

DO YOU WANT AN ANOVA TABLE(Y/N)?- Y

ANOVA TABLE
  SOURCE        DF          SS              MS              F
REGRESSION      2.    0.265786E+04    0.132893E+04    229.5037
ERROR          10.    0.579045E+02    0.579045E+01
TOTAL          12.    0.271576E+04

DO YOU WANT A TABLE OF RESIDUALS(Y/N)?-
```

```
DO YOU WISH ANOTHER RUN WITH THIS FILE(Y/N)?-

TYPE:NO OF THE DEP VAR(Y), NO OF INDEP VARS(X)- 5,3

TYPE:  3 NOS (IN ORDER) FOR INDEP VARS, MAX 20 PER LINE-
1,2,3

FOR RIDGE REGRESSION, TYPE A VALUE BETWEEN 0 & 1-


 7-MAY-80 FILE THALD
HALD DATA GIVEN BY DRAPER & SMITH (1966), P.178
X#      MEAN          ST DEV        R(XY)          REG CO       T VALUE
 1 0.74615E+01 0.58824E+01 0.73072E+00 0.16958BE+01   8.2895
 2 0.48154E+02 0.15561E+02 0.81625E+00 0.65691SE+00  14.8508
 3 0.11769E+02 0.64051E+01-0.53467E+00 0.25001BE+00   1.3536
Y#
 5 0.95423E+02 0.15044E+02                0.48193BE+02

R=.99110 RSQ=.98228 SEE=.23121E+01 ERRMS=.53456E+01 ERRSS=.48111E+02

DO YOU WANT AN ANOVA TABLE(Y/N)?- Y

ANOVA TABLE
  SOURCE          DF          SS                MS              F
REGRESSION        3.    0.266765E+04    0.889217E+03    166.3449
ERROR             9.    0.481106E+02    0.534562E+01
TOTAL            12.    0.271576E+04

DO YOU WANT A TABLE OF RESIDUALS(Y/N)?-
```

```
DO YOU WISH ANOTHER RUN WITH THIS FILE(Y/N)?-

TYPE:NO OF THE DEP VAR(Y), NO OF INDEP VARS(X)- 5,3

TYPE:  3 NOS (IN ORDER) FOR INDEP VARS, MAX 20 PER LINE-
1,2,4

FOR RIDGE REGRESSION, TYPE A VALUE BETWEEN 0 & 1-


 7-MAY-80 FILE THALD
HALD DATA GIVEN BY DRAPER & SMITH (1966), P.178
X#      MEAN         ST DEV       R(XY)         REG CO       T VALUE
 1 0.74615E+01 0.58824E+01 0.73072E+00 0.14519AE+01   12.4100
 2 0.48154E+02 0.15561E+02 0.81625E+00 0.41611BE+00    2.2418
 4 0.30000E+02 0.16738E+02-0.82131E+00-0.23654BE+00   -1.3650
Y#
 5 0.95423E+02 0.15044E+02                0.71648BE+02

R=.99113 RSQ=.98234 SEE=.23087E+01 ERRMS=.53303E+01 ERRSS=.47973E+02

DO YOU WANT AN ANOVA TABLE(Y/N)?- Y

ANOVA TABLE
  SOURCE        DF          SS              MS              F
REGRESSION      3.    0.266779E+04    0.889263E+03    166.8317
ERROR           9.    0.479727E+02    0.533030E+01
TOTAL          12.    0.271576E+04

DO YOU WANT A TABLE OF RESIDUALS(Y/N)?-
```

```
DO YOU WISH ANOTHER RUN WITH THIS FILE(Y/N)?-


TYPE:NO OF THE DEP VAR(Y), NO OF INDEP VARS(X)- 5,4

TYPE:  4 NOS (IN ORDER) FOR INDEP VARS, MAX 20 PER LINE-
1,2,3,4

FOR RIDGE REGRESSION, TYPE A VALUE BETWEEN 0 & 1-



 7-MAY-80 FILE THALD
HALD DATA GIVEN BY DRAPER & SMITH (1966), P.178
X#      MEAN         ST DEV       R(XY)         REG CO       T VALUE
 1 0.74615E+01 0.58824E+01 0.73072E+00 0.15511QE+01   2.0827
 2 0.48154E+02 0.15561E+02 0.81625E+00 0.51016RE+00   0.7049
 3 0.11769E+02 0.64051E+01-0.53467E+00 0.10190QE+00   0.1350
 4 0.30000E+02 0.16738E+02-0.82131E+00-0.14406IE+00  -0.2032
Y#
 5 0.95423E+02 0.15044E+02                0.62405AE+02

R=.99115 RSQ=.98238 SEE=.24460E+01 FRMS=.59830E+01 FRSS=.47864E+02

DO YOU WANT AN ANOVA TABLE(Y/N)?- Y

ANOVA TABLE
  SOURCE        DF         SS              MS              F
REGRESSION      4.    0.266790E+04   0.666975E+03    111.4792
ERROR           8.    0.478636E+02   0.598295E+01
TOTAL          12.    0.271576E+04

DO YOU WANT A TABLE OF RESIDUALS(Y/N)?- Y

OBS#    OBSERVED Y   ESTIMATED Y    OBSY-ESTY   %(O-E)/E
   1  0.78500E+02  0.78495E+02  0.47604E-02    0.006
   2  0.74300E+02  0.72789E+02  0.15112E+01    2.076
   3  0.10430E+03  0.10597E+03 -0.16709E+01   -1.577
   4  0.87600E+02  0.89327E+02 -0.17271E+01   -1.933
   5  0.95900E+02  0.95649E+02  0.25076E+00    0.262
   6  0.10920E+03  0.10527E+03  0.39254E+01    3.729
   7  0.10270E+03  0.10415E+03 -0.14487E+01   -1.391
   8  0.72500E+02  0.75675E+02 -0.31750E+01   -4.196
   9  0.93100E+02  0.91722E+02  0.13783E+01    1.503
  10  0.11590E+03  0.11562E+03  0.28155E+00    0.244
  11  0.83800E+02  0.81809E+02  0.19910E+01    2.434
  12  0.11330E+03  0.11233E+03  0.97299E+00    0.866
  13  0.10940E+03  0.11169E+03 -0.22943E+01   -2.054
```

```
DO YOU WISH ANOTHER RUN WITH THIS FILE(Y/N)?-


TYPE:NO OF THE DEP VAR(Y), NO OF INDEP VARS(X)- 5,4

TYPE:  4 NOS (IN ORDER) FOR INDEP VARS, MAX 20 PER LINE-
1,2,3,4

FOR RIDGE REGRESSION, TYPE A VALUE BETWEEN 0 & 1- .03



 7-MAY-80 FILE THALD
HALD DATA GIVEN BY DRAPER & SMITH (1966), P.178
X#      MEAN         ST DEV       R(XY)         REG CO       T VALUE
 1 0.74615E+01 0.58824E+01 0.73072E+00 0.124134E+01   2.6672
 2 0.48154E+02 0.15561E+02 0.81625E+00 0.289723E+00   0.9115
 3 0.11769E+02 0.64051E+01-0.53467E+00-0.185178E+00  -0.4210
 4 0.30000E+02 0.16738E+02-0.82131E+00-0.356245E+00  -1.1564
Y#
 5 0.95423E+02 0.15044E+02                0.850762E+02

R=.96893 RSQ=.93883 SEE=.46247E+01 ERRMS=.21388E+02 ERRSS=.17110E+03

DO YOU WANT AN ANOVA TABLE(Y/N)?- Y

ANOVA TABLE
  SOURCE         DF          SS                MS                F
REGRESSION       4.    0.262613E+04    0.656533E+03     30.6964
ERROR            8.    0.171104E+03    0.213880E+02
TOTAL           12.    0.279724E+04

DO YOU WANT A TABLE OF RESIDUALS(Y/N)?- Y

OBS#    OBSERVED Y   ESTIMATED Y    OBSY-ESTY   %(O-E)/E
   1  0.78500E+02  0.78813E+02 -0.31266E+00   -0.397
   2  0.74300E+02  0.73417E+02  0.88288E+00    1.203
   3  0.10430E+03  0.10635E+03 -0.20492E+01   -1.927
   4  0.87600E+02  0.89487E+02 -0.18875E+01   -2.109
   5  0.95900E+02  0.95964E+02 -0.64070E-01   -0.067
   6  0.10920E+03  0.10516E+03  0.40382E+01    3.840
   7  0.10270E+03  0.10409E+03 -0.13851E+01   -1.331
   8  0.72500E+02  0.75550E+02 -0.30503E+01   -4.037
   9  0.93100E+02  0.92033E+02  0.10667E+01    1.159
  10  0.11590E+03  0.11476E+03  0.11417E+01    0.995
  11  0.83800E+02  0.81535E+02  0.22649E+01    2.778
  12  0.11330E+03  0.11191E+03  0.13888E+01    1.241
  13  0.10940E+03  0.11143E+03 -0.20345E+01   -1.826

IN ORIGINAL UNITS: SEE=.24953E+01 ERRMS=.62265E+01 ERRSS=.49812E+02
```

```
DO YOU WISH ANOTHER RUN WITH THIS FILE(Y/N)?-

TYPE:NO OF THE DEP VAR(Y), NO OF INDEP VARS(X)- 5,4

TYPE:  4 NOS (IN ORDER) FOR INDEP VARS, MAX 20 PER LINE-
1,2,3,4

FOR RIDGE REGRESSION, TYPE A VALUE BETWEEN 0 & 1- .002


 7-MAY-80 FILE THALD
HALD DATA GIVEN BY DRAPER & SMITH (1966), P.178
X#      MEAN          ST DEV        R(XY)          REG CO       T VALUE
 1 0.74615E+01 0.58824E+01 0.73072E+00 0.141283E+01   2.5039
 2 0.48154E+02 0.15561E+02 0.81625E+00 0.377947E+00   0.7167
 3 0.11769E+02 0.64051E+01-0.53467E+00-0.374105E-01  -0.0661
 4 0.30000E+02 0.16738E+02-0.82131E+00-0.273463E+00  -0.5298
Y#
 5 0.95423E+02 0.15044E+02                0.753258E+02

R=.98956 RSQ=.97924 SEE=.26575E+01 ERRMS=.70621E+01 ERRSS=.56497E+02

DO YOU WANT AN ANOVA TABLE(Y/N)?- Y

ANOVA TABLE
  SOURCE        DF          SS              MS              F
REGRESSION      4.    0.266470E+04    0.666174E+03     94.3310
ERROR           8.    0.564967E+02    0.706209E+01
TOTAL          12.    0.272119E+04

DO YOU WANT A TABLE OF RESIDUALS(Y/N)?- Y

OBS#    OBSERVED Y   ESTIMATED Y    OBSY-ESTY   %(O-E)/E
   1  0.78500E+02  0.78410E+02  0.90036E-01    0.115
   2  0.74300E+02  0.72918E+02  0.13822E+01    1.896
   3  0.10430E+03  0.10626E+03 -0.19634E+01   -1.848
   4  0.87600E+02  0.89431E+02 -0.18312E+01   -2.048
   5  0.95900E+02  0.95620E+02  0.27992E+00    0.293
   6  0.10920E+03  0.10530E+03  0.38989E+01    3.703
   7  0.10270E+03  0.10412E+03 -0.14217E+01   -1.365
   8  0.72500E+02  0.75600E+02 -0.30996E+01   -4.100
   9  0.93100E+02  0.91871E+02  0.12290E+01    1.338
  10  0.11590E+03  0.11550E+03  0.40097E+00    0.347
  11  0.83800E+02  0.81698E+02  0.21017E+01    2.573
  12  0.11330E+03  0.11219E+03  0.11068E+01    0.987
  13  0.10940E+03  0.11157E+03 -0.21736E+01   -1.948

IN ORIGINAL UNITS: SEE=.24514E+01 ERRMS=.60092E+01 ERRSS=.48073E+02
```